

Evidenz, Validierung und Entscheidung durch Prüfungsgremien

Entscheidungsziel: Bestimmen, welche Behauptung durch welche Nachweise gedeckt ist und welche begrenzte Entscheidung daraus folgt.

Fassung 1.0 · 25. September 2026 · Adressaten: Wissenschaftliche und technische Prüfungsgremien, Qualitätsverantwortliche, Evaluatoren, Sicherheits- und Beschaffungsausschüsse, Forschungsleitungen und Fachredaktionen.

Dokumenttyp: strukturierte integrative Recherche mit abgeleiteten Verfahren, keine systematische Übersichtsarbeit, keine eigene empirische Robotikstudie und keine einzelfallbezogene Rechts- oder Förderberatung.

Eine unabhängige Begutachtung, Zertifizierung oder fallbezogene Rechtsberatung ist mit der Erstellung dieses Leitfadens nicht erfolgt.

Steckbrief

Merkmal	Wert
Dokumenttyp	Strukturierte integrative Recherche mit abgeleiteten Verfahren; keine systematische Übersichtsarbeit, keine unabhängige Fachbegutachtung
Zielgruppe	Wissenschaftliche und technische Prüfungsgremien, Qualitätsverantwortliche, Evaluatoren, Sicherheits- und Beschaffungsausschüsse, Forschungsleitungen und Fachredaktionen.
Recherche- und Rechtsstand	25. September 2026
Umfang	3.258 Wörter, 13 Seiten
Belege	0 Literatureinträge, 4 Quellenadressen
Gliederung	14 nummerierte Abschnitte, 0 Tabellen

Inhalt

1. Evidenzprüfung beginnt vor der Statistik	5
2. Die Quellenrolle und die Nachweisdimensionen bestimmen	5
2.1 Keine lineare Leiter unterschiedlicher Eigenschaften	5
2.2 Ein konkreter Quellenkonflikt als Lehrfall	5
3. Ein Protokoll vor Datenerhebung festlegen	5
4. Versuchseinheit, Zielpopulation und Auswahlverzerrung	6
4.1 Angeforderte, zugelassene und gestartete Aufgaben	6
4.2 Warum mehr Aufgaben nicht immer mehr unabhängige Evidenz bedeuten	6
5. Vergleichsdesign und kausale Interpretation	6
6. Primäre Ergebnisse, Qualitätsprüfung und Interventionsbuchhaltung	7
6.1 Ein synthetisches, vollständig auflösbares Beispiel	7
6.2 Zustände und Ursachen auseinanderhalten	7
6.3 Messfehler und Beobachterübereinstimmung	7
7. Was null beobachtete Fehler statistisch bedeuten	7
7.1 Binomialmodell mit expliziten Annahmen	7
7.2 Welche Stichprobe wäre für eine engere Behauptung nötig?	8
7.3 Mehr als null Fehler	8
8. Zeitbezogene Zuverlässigkeit und ihre Grenzen	8
9. Warum ein hoher Gesamtwert die falsche Auswahl begründen kann	9
10. Generalisierung, Standortwechsel und Modellversionen	9
11. Fehlende Daten, selektive Veröffentlichung und Widerlegungsversuche	9
12. Von Messdaten zu einem belastbaren Nachweisargument	9
13. Verfahren eines fachlichen Prüfgremiums	10
13.1 Rollen und Vorbereitung	10
13.2 Beschluss und Nachverfolgung	10
14. Reproduzierbarkeit des Prüfprozesses	10
Quellen zu diesem Dokument	11
Q009 · Robotics Institute Germany	11
Q015 · Figure AI	11
Q016 · BMW Group	11
Q018 · NIST/SEMATECH	11

12 Rechtliche Hinweise **12**

12.1 Keine Beratung	12
12.2 Keine Zusicherung, keine Konformitätsaussage	12
12.3 Rechtsstand	12
12.4 Haftung	12
12.5 Verweise auf fremde Inhalte	12
12.6 Marken und Produktbezeichnungen	12

12.7 Urheberrecht und Nutzung	12
12.8 Keine personenbezogenen Daten	13
12.9 Unabhängigkeit	13
12.10 Anwendbares Recht	13

Impressum und Verantwortlichkeit

Die folgenden Angaben sind wörtlich dem Impressum von fluxlane.eu entnommen, abgerufen am 24. September 2026. Sie wurden nicht ergänzt und nicht umformuliert.

Angabe	Inhalt
Angaben gemäß § 5 DDG	WebTrades GmbH Am Ölberg 2 55459 Grolsheim Deutschland
Vertreten durch	Geschäftsführerin: Laurice Bussmer
Kontakt	Telefon: 0170 890 78 00 E-Mail: info@fluxlane.eu
Registereintrag	Eintragung im Handelsregister Registergericht: Amtsgericht Mainz Registernummer: HRB 54479
Umsatzsteuer-ID	Umsatzsteuer-Identifikationsnummer gemäß § 27 a Umsatzsteuergesetz: DE461650292
Redaktionell verantwortlich	Verantwortlich für den Inhalt nach § 18 Abs. 2 MStV: Laurice Bussmer, Am Ölberg 2, 55459 Grolsheim
Verbraucherstreitbeilegung	Wir sind nicht bereit oder verpflichtet, an Streitbeilegungsverfahren vor einer Verbraucherschlichtungsstelle teilzunehmen.

Redaktioneller Hinweis des Herausgebers (wörtlich)

„fluxlane ist ein unabhängiges, redaktionelles Fachportal für Robotik, autonome physische Systeme und verkörperte KI. Der redaktionelle Geltungsbereich umfasst Industrie und Produktion, Logistik, Medizin, Pflege und Rehabilitation, Service und Haushalt sowie Forschung, Hersteller, Markt und Regulierung; künstliche Intelligenz gehört dazu, auch ohne physischen Körper. Die Inhalte dienen der allgemeinen Information und stellen keine Rechts-, Kauf- oder Anlageberatung dar. Angaben zu Preisen, Verfügbarkeit und Terminen erfolgen ohne Gewähr und können sich jederzeit ändern.“

Zitierweise

FluxLane (2026): *Evidenz, Validierung und Entscheidung durch Prüfungsgremien*. Fassung 1.0, Stand 25. September 2026. Herausgegeben von der WebTrades GmbH, Grolsheim, fluxlane.eu.

Herstellung dieses Dokuments

Das vorliegende PDF ist maschinell aus der Markdown-Fassung des Leitfaden-Pakets gesetzt. Inhaltlich ist dieser Satzvorgang keine neue Sachprüfung; die zugrunde liegende Recherche- und Quellenprüfung ist im Lieferpaket dokumentiert. Gesetzt wurde mit IBM Plex Sans und Sora.

Abbildungsnachweis

Dieses Dokument enthält keine Abbildungen.

1. Evidenzprüfung beginnt vor der Statistik

Eine exakte Berechnung kann eine unklare Frage nicht reparieren. Bevor Erfolgsquoten, Konfidenzgrenzen oder Betriebsstunden bewertet werden, muss der Gegenstand der Behauptung feststehen. Gemeint sein können eine einzelne Handlung, ein vollständiger Auftrag, eine Schicht, eine Maschine, ein Standort oder eine ganze Flotte. Diese Einheiten sind nicht austauschbar. Auch die interessierende Eigenschaft muss definiert sein: physischer Erfolg, qualitätsgerechtes Ergebnis, autonome Ausführung, betriebliche Verfügbarkeit oder wirtschaftlicher Nutzen.

Der erste Prüfschritt formuliert daher einen **Zielparameter**. Ein Beispiel ist der Anteil aller innerhalb des freigegebenen Einsatzkorridors angeforderten Aufträge, die eine bestimmte Konfiguration innerhalb der zugesagten Zeit qualitätsgerecht und ohne menschlichen Eingriff beendet. Ein anderer Zielparameter wäre die Zahl menschlicher Interventionereignisse je 1.000 aktive Betriebsstunden. Beide sind legitim, beantworten aber unterschiedliche Fragen.

Ein gutes Prüfprogramm prüft außerdem, ob die ausgewählte Messgröße das behauptete Konstrukt tatsächlich abbildet. Eine interne „success“-Meldung misst zunächst die Selbstbewertung des Systems. Sie ist nicht identisch mit einem unabhängig geprüften Ergebnis. Ebenso kann ein hoher Durchsatz außerhalb der erforderlichen Qualität liegen. Diese Fragen der Messgültigkeit gehen jeder präzisen Statistik voraus.

2. Die Quellenrolle und die Nachweisdimensionen bestimmen

2.1 Keine lineare Leiter unterschiedlicher Eigenschaften

Evidenz wird in diesem Paket entlang mehrerer Achsen beschrieben: Identifizierbarkeit des Systems, Reproduzierbarkeit, Realitätsnähe, Dauer, Vielfalt der Bedingungen, Unabhängigkeit und Qualität der Messdefinitionen. Die Achsen werden nicht zu einem Gesamtwert addiert. Ein großer Flottenbestand ersetzt keinen unabhängigen Nachweis; eine unabhängige Laborprüfung ersetzt keinen repräsentativen Dauerbetrieb.

Die Unabhängigkeit selbst ist graduell und sachbezogen zu beschreiben. Ein Kundenbericht kann weniger unmittelbar vom Hersteller stammen, aber aus einer gemeinsamen Entwicklungspartnerschaft resultieren. Eine externe Prüfstelle kann unabhängig sein, jedoch nur eine enge elektrische oder mechanische Teilfrage untersucht haben. Die Aussage „extern geprüft“ muss deshalb immer den Prüfgegenstand, Auftraggeber, Standard beziehungsweise Verfahren und Geltungsbereich nennen.

Für fachjournalistische Einordnung genügt häufig ein genau abgegrenzter öffentlicher Beleg. Für eine betriebliche Freigabe werden regelmäßig weitere interne Dokumente benötigt. Der Unterschied ist kein Widerspruch: Die zulässige Schlussfolgerung richtet sich nach der Entscheidung, nicht allein nach dem Publikationsstatus der Quelle.

2.2 Ein konkreter Quellenkonflikt als Lehrfall

Die Hersteller- und Betreiberberichte zu Figure 02 in Spartanburg sind für die Existenz eines quantifizierten Einsatzes aussagekräftiger als ein Demonstrationsvideo. Sie nennen jedoch unterschiedliche Zeitspannen. Figure bezeichnet zudem null Interventionen pro Schicht als Ziel, ohne damit eine vollständig veröffentlichte Ist-Verteilung zu liefern. Ein Prüfer hält deshalb getrennt fest, was berichtet wurde und welche weitergehenden Größen offen bleiben. [Q015, Deployment und Core KPIs; Q016, Spartanburg-Rückblick]

Aus solchen Quellen kann nicht die Wahrscheinlichkeit eines unfallfreien zukünftigen Betriebs berechnet werden. Es fehlen unter anderem vollständig definierte Exposition, Ereignistaxonomie und repräsentative Daten. Die richtige Konsequenz ist eine begrenzte Aussage über öffentlich verfügbare Betriebsnachweise, nicht eine erfundene Zuverlässigkeitskennzahl.

3. Ein Protokoll vor Datenerhebung festlegen

Ein auswertbares Protokoll enthält Forschungs- oder Prüffrage, Zielpopulation, eingeschlossene Aufgaben, Ausschlüsse, Konfiguration, Vergleich, Messverfahren, geplante Auswertung und Abbruchregeln. Es legt fest, wann ein Versuch beginnt und endet, wie Wiederholungen behandelt werden und welche Unterstützung zulässig ist. Änderungen werden datiert und begründet; die ursprüngliche Fassung bleibt erhalten.

Die Vorabfestlegung schützt vor ergebnisabhängiger Umdefinition. Ein Projekt darf seinen Entwicklungsplan aufgrund neuer Erkenntnisse ändern. Es darf anschließend jedoch nicht so tun, als sei der geänderte Test schon immer die ursprüngliche Bestätigungshypothese gewesen. Explorative Analyse und bestätigende Prüfung werden getrennt berichtet. Das ist insbesondere relevant, wenn viele Objektklassen, Metriken oder Parameter ausprobiert werden und später nur die günstigste Kombination dargestellt wird.

Ein Pilot benötigt außerdem eine dokumentierte Sicherheits- und Rechtsfreigabe für seine konkrete Versuchsanordnung. Statistische Repräsentativität rechtfertigt keine unzulässige Exposition von Personen. Gefährliche Randbedingungen können zunächst durch Analyse, Simulation oder geeignete geschützte Versuchsanordnungen untersucht werden. Die Prüfmethodik entscheidet nicht eigenmächtig, welche physischen Risiken akzeptiert werden dürfen.

4. Versuchseinheit, Zielpopulation und Auswahlverzerrung

4.1 Angeforderte, zugelassene und gestartete Aufgaben

Ein System kann eine hohe Erfolgsquote erreichen, indem es schwierige Aufgaben gar nicht annimmt. Das ist innerhalb eines klar definierten Einsatzkorridors möglicherweise sinnvoll. Für die betriebliche Eignung sind dann jedoch zwei Größen erforderlich: der Anteil der betrieblich benötigten Aufgaben, die das System übernehmen darf und kann, sowie der Erfolg innerhalb dieser angenommenen Aufgaben. Werden beide vermischt, wird ein enges Spezialverfahren als umfassende Lösung dargestellt.

Die Ereigniskette sollte daher Anfrage, Zulassung, Start, Unterbrechung, Endzustand und Qualitätsentscheidung getrennt erfassen. Nicht zugelassene Aufgaben bleiben mit Grund im Register. Unzulässige oder außerhalb des Einsatzkorridors liegende Aufgaben werden nicht künstlich als technische Fehler gewertet; sie bleiben jedoch für die Frage relevant, welchen Anteil des tatsächlichen Bedarfs die Lösung abdeckt.

Auch Ausfälle vor dem ersten Aufgabenstart sind relevant. Eine Erfolgsquote aus begonnenen Aufgaben kann gut aussehen, obwohl das System oft nicht einsatzbereit war. Die zeitliche Verfügbarkeit und die auftragsbezogene Leistung werden deshalb separat berichtet.

4.2 Warum mehr Aufgaben nicht immer mehr unabhängige Evidenz bedeuten

Wiederholungen am selben Roboter unter nahezu gleichen Bedingungen teilen gemeinsame Einflüsse. Ein fehlerhafter Sensor, eine besondere Beleuchtung oder ein spezifischer Bediener kann viele Beobachtungen gleichzeitig beeinflussen. Für Aussagen über neue Roboter oder Standorte ist die Zahl der unabhängigen Einheiten daher wichtiger als die reine Zahl einzelner Aktionen.

Ein einfaches **Illustrationsmodell**, nicht eine allgemeine Stichprobenformel, ist der Designeffekt $D = 1 + (m-1)p$ bei gleich großen Clustern der Größe m und angenommener Intracuster-Korrelation p . Bei $m = 100$ und $p = 0,05$ ergibt sich $D = 5,95$. Aus 1.000 Beobachtungen würde in dieser groben Näherung eine effektive Information entsprechend etwa 168 unabhängigen Beobachtungen. Diese Näherung darf nicht ungeprüft in einen exakten Zulassungsnachweis eingesetzt werden. Sie zeigt, warum Clusterstruktur bei der Planung berücksichtigt werden muss.

Für reale Programme kommen je nach Datenlage gemischte Modelle, clusterrobuste Auswertung oder auf der unabhängigen Einheit beruhende Resampling-Verfahren in Betracht. Die Wahl muss vor der abschließenden Auswertung begründet und durch statistische Fachkunde abgesichert werden. Bei wenigen Standorten lässt sich eine breite Populationsaussage auch durch komplexe Modelle nicht beliebig herstellen.

5. Vergleichsdesign und kausale Interpretation

Ein Vorher-nachher-Vergleich kann durch Produktmix, Saison, Personal, Lernkurve oder gleichzeitige Prozessänderungen verzerrt sein. Wenn während des Roboterpilots Material besser vorbereitet wird, kann ein beobachteter Leistungsgewinn auf dieser Vorbereitung beruhen. Das ist betrieblich möglicherweise erwünscht, aber nicht ausschließlich eine Wirkung des Roboters.

Wo sicher, rechtlich zulässig und praktisch möglich, können vergleichbare Aufgaben oder Zeitblöcke nach einem vorab festgelegten Verfahren unterschiedlichen Varianten zugeteilt werden. Blockierung nach Schicht oder Objektfamilie kann wichtige Unterschiede kontrollieren. Bei organisatorischen Lern- und Übertragungseffekten sind Einzelaufträge jedoch möglicherweise keine unabhängige Vergleichseinheit; dann sind größere Zeit- oder Standortblöcke zu prüfen. Eine bloße Randomisierung beseitigt nicht jede Interferenz zwischen den Bedingungen.

Ein nicht randomisiertes Design muss seine Annahmen besonders offenlegen. Eine Vergleichsgruppe ist nur dann hilfreich, wenn sie den relevanten unbeobachteten Verlauf ausreichend repräsentiert. Verfahren wie ein Differenz-von-Differenzen-Vergleich benötigen zusätzliche Voraussetzungen; der Name eines statistischen Verfahrens ersetzt deren Prüfung nicht. Wo die Daten nur eine Beschreibung erlauben, sollte der Bericht genau dabei bleiben.

6. Primäre Ergebnisse, Qualitätsprüfung und Interventionsbuchhaltung

6.1 Ein synthetisches, vollständig auflösbares Beispiel

Im beigefügten Datensatz werden 1.000 Aufgaben dargestellt. Es handelt sich ausdrücklich um künstlich erzeugte Demonstrationsdaten ohne Zufallsstichprobe aus einem realen Betrieb. Die Kategorien sind disjunkt: 970 Erstversuchserfolge, 18 autonome Recoveries, zehn nach menschlicher Hilfe akzeptierte Ergebnisse und zwei ungelöste Aufgaben. Daraus folgen:

$$\text{Erstversuchserfolg} = 970 / 1000 = 97,0 \%$$

$$\text{Autonome Enderledigung} = (970 + 18) / 1000 = 98,8 \%$$

$$\text{Schließlich akzeptierte Ergebnisse} = (970 + 18 + 10) / 1000 = 99,8 \%$$

Die zwölf Aufgaben mit menschlicher Hilfe enthalten zusammen 14 Interventionsereignisse. Die auftragsbezogene Hilfsquote beträgt daher 1,2 Prozent; die Ereignisrate beträgt 1,4 Interventionen je 100 gestartete Aufgaben. Das sind unterschiedliche Nenner. Eine Intervention kann zudem sehr kurz oder sehr lang sein. Ihre Dauer und erforderliche Qualifikation müssen für Wirtschaftlichkeit und Betriebsfähigkeit zusätzlich betrachtet werden.

6.2 Zustände und Ursachen auseinanderhalten

Ein sicherheitsbedingter Stopp kann die Nutzfunktion unterbrechen und zugleich zeigen, dass eine Schutzfunktion wirksam reagiert hat. Er ist deshalb weder pauschal „Erfolg“ noch pauschal „Versagen“. Im Register werden Anlass, Schutzreaktion, Betriebsfolge und Ergebnis getrennt festgehalten. Entsprechend kann ein Qualitätsfehler ohne Sicherheitsvorfall wirtschaftlich relevant sein.

Ein gutes Ereignisschema erlaubt mehrere Folgen, aber verhindert Doppelzählung des gleichen primären Ereignisses. Es dokumentiert auch die Auflösung: autonom behoben, durch Bedienperson behoben, durch Service behoben oder ungelöst. „Reset erfolgreich“ genügt nicht, wenn die ursprüngliche Aufgabe verloren ging oder anschließend doppelt ausgeführt wurde.

6.3 Messfehler und Beobachterübereinstimmung

Die Qualitätsreferenz muss beschrieben werden. Bei visueller oder manueller Beurteilung sind Kriterien, Schulung und Umgang mit Zweifelsfällen festzulegen. Eine Stichprobe kann von zwei Personen unabhängig bewertet werden, um Uneinigkeit sichtbar zu machen. Der bloße Prozentsatz übereinstimmender Bewertungen genügt nicht immer, wenn eine Klasse sehr selten ist. Entscheidend bleibt, welche Fehlentscheidungen für die Anwendung relevant sind.

Wenn die automatische Messpipeline dieselben Komponenten wie das Prüfsystem nutzt, sind gemeinsame Fehler möglich. Daher ist zu prüfen, welche Ergebnisse über unabhängige Daten oder eine geeignete Referenz bestätigt werden müssen. Nachträgliche Korrekturen der Labels werden versioniert; sonst kann eine scheinbare Modellverbesserung aus veränderter Bewertungspraxis stammen.

7. Was null beobachtete Fehler statistisch bedeuten

7.1 Binomialmodell mit expliziten Annahmen

Angenommen, n Versuche sind unabhängig, besitzen dieselbe unbekannte Fehlerwahrscheinlichkeit p und wurden vollständig beobachtet. Dann ist die Wahrscheinlichkeit, keinen Fehler zu sehen, $(1-p)^n$. Für null beobachtete Fehler ergibt sich die exakte einseitige obere Konfidenzgrenze zum Niveau $1-\alpha$ aus:

$$p_{\text{oben}} = 1 - \alpha^{(1/n)}.$$

Bei $n = 1.000$ und $\alpha = 0,05$ ist $p_{\text{oben}} = 0,00299125$, also rund **0,2991 Prozent**. Null beobachtete Fehler bedeuten in diesem Modell somit nicht null Fehlerwahrscheinlichkeit. Die häufige Näherung $3/n$ ergibt für ein 95-Prozent-Niveau einen ähnlichen Wert, weil $-\ln(0,05)$ ungefähr drei beträgt. Diese Herleitung ist elementare eigene Mathematik; sie ersetzt nicht die Prüfung der Modellannahmen.

Die frequentistische Aussage lautet nicht, dass der feste unbekannte Parameter mit 95 Prozent Wahrscheinlichkeit in diesem konkreten Intervall liegt. Gemeint ist die Abdeckung des verwendeten Intervallverfahrens bei wiederholter Anwendung unter seinen Voraussetzungen. Diese Unterscheidung ist wichtig, sobald Gremien Konfidenz mit subjektiver Sicherheit verwechseln.

7.2 Welche Stichprobe wäre für eine engere Behauptung nötig?

Soll bei null Fehlern eine obere Grenze von höchstens 0,1 Prozent erreicht werden, muss gelten:

$$n \geq \ln(0,05) / \ln(1-0,001).$$

Das ergibt aufgerundet **2.995 unabhängige gleichartig verteilte Versuche**. Diese Zahl ist kein allgemeiner Robotik-Mindesttest. Sie hängt am gewählten Ziel, Konfidenzniveau und Modell. Sie behandelt weder unterschiedliche Fehlerfolgen noch Cluster, veränderte Software oder unbekannte Umgebungen. Ein Sicherheitsnachweis wird nicht durch das bloße Abarbeiten dieser Anzahl ersetzt.

Wenn während der Durchführung wiederholt auf die Daten geschaut und beim ersten günstigen Ergebnis beendet wird, kann die nominelle Aussagekraft des ursprünglich geplanten Tests verloren gehen. Ein sequenzielles Verfahren ist möglich, benötigt aber eine dafür geeignete vorab definierte Auswertung. „Wir testen, bis wir bestehen“ ist kein neutraler Stichprobenplan.

7.3 Mehr als null Fehler

Bei k beobachteten Fehlern kann eine einseitige exakte obere Grenze durch Inversion der Binomialverteilung bestimmt werden: Gesucht wird das p , für das $P_p(X \leq k) = \alpha$ gilt. Das mitgelieferte Python-Skript implementiert diese Inversion ohne zusätzliche Statistikbibliothek für die demonstrierten Fallgrößen. Es dokumentiert auch Randfälle. Die mathematisch korrekte Inversion macht eine selektiv gewählte oder abhängige Stichprobe jedoch nicht repräsentativ.

8. Zeitbezogene Zuverlässigkeit und ihre Grenzen

Für reparierbare Systeme kann ein homogener Poisson-Prozess ein einfaches Modell von Ereignissen über die Zeit sein. Die Ereignisrate wird durch Ereigniszahl geteilt durch Exposition geschätzt. Das NIST-Handbuch beschreibt dieses Modell und seine Konfidenzverfahren ausdrücklich im Kontext einer konstanten Rate. [Q018, Abschnitt 8.4.5.1]

Bei null Ereignissen in einer vollständig beobachteten Zeit T ergibt sich in diesem Modell die einseitige obere Rate $\lambda_{\text{oben}} = -\ln(\alpha)/T$. Für ein **rein hypothetisches** Beispiel mit $T = 2.000$ Stunden und $\alpha = 0,05$ sind das rund 0,00149787 Ereignisse pro Stunde oder 1,4979 Ereignisse pro 1.000 Stunden. Es handelt sich nicht um eine Aussage über Figure, BMW oder einen anderen realen Anbieter.

Die Verwendung von Betriebsstunden als Exposition erfordert eine Definition. Zählt eingeschaltete Zeit, aktive Aufgabenbearbeitung, verfügbare Zeit oder Zeit mit potenzieller Gefährdung? Unterschiedliche Fragen benötigen unterschiedliche Nenner. Eine Akkumulation von Labor- und Kundenstunden kann beschreibend sinnvoll sein, ist aber ohne Trennung nur eingeschränkt für eine bestimmte Nutzungsbehauptung geeignet.

Konstante Raten sind bei lernenden, alternden oder häufig aktualisierten Systemen keineswegs selbstverständlich. Frühe Anlauffehler, Verschleiß, saisonale Einflüsse oder geänderte Modelle können die Rate verändern. Die

Auswertung sollte deshalb Zeitverlauf, Roboteridentität und Konfiguration berücksichtigen. Ein einziger MTBF- oder MTBI-Wert über alle Generationen kann eine vermeintliche Präzision erzeugen, die keine konkrete Version beschreibt.

9. Warum ein hoher Gesamtwert die falsche Auswahl begründen kann

Ein synthetisches Beispiel verdeutlicht die Bedeutung des Aufgabenmixes. System A erledigt 900 einfache Aufgaben mit 99 Prozent Erfolg und 100 schwierige mit 80 Prozent. Insgesamt ergeben sich 971 Erfolge aus 1.000 Versuchen, also 97,1 Prozent. System B erledigt 100 einfache Aufgaben mit 99 Prozent und 900 schwierige mit 90 Prozent. Insgesamt sind es 909 von 1.000, also 90,9 Prozent.

Die aggregierten Werte lassen A überlegen erscheinen, obwohl B in der einfachen Gruppe gleich gut und in der schwierigen besser ist. Bei einem Zielmix von je 50 Prozent lägen die standardisierten Erfolgswerte bei 89,5 Prozent für A und 94,5 Prozent für B. Das Beispiel illustriert eine Mischungsverzerrung. Es ist keine statistische Signifikanzprüfung und keine reale Produktbewertung.

Für die Praxis folgt: Ergebnisse werden nach vorab relevanten Gruppen dargestellt und auf den vorgesehenen Einsatzmix bezogen. Eine Standardisierung kann helfen, benötigt aber begründete Gewichte. Werden Gewichte nachträglich so gewählt, dass ein bevorzugtes System gewinnt, entsteht eine andere Form der Verzerrung. Die Zielpopulation gehört daher in das Prüfprotokoll.

10. Generalisierung, Standortwechsel und Modellversionen

Generalisierung ist eine Behauptung über nicht vollständig beobachtete Bedingungen. Sie kann sich auf neue Objekte, Umgebungen, Nutzer oder Aufgaben beziehen. Jede dieser Varianten benötigt ein geeignetes Design. Ein neuer Hintergrund im Bild ist nicht gleichbedeutend mit einem neuen physischen Prozess. Ein neuer Standort mit identischem Layout ist nicht dasselbe wie ein neuer Aufgabentyp.

Ein Nachweis sollte beschreiben, welche Dimensionen variiert wurden und welche konstant blieben. Für eine geplante Übertragung wird anschließend geprüft, ob die Unterschiede innerhalb dieses untersuchten Bereichs liegen. Wo dies nicht der Fall ist, bleiben zusätzliche Anforderungen offen. Das lässt sich als Varianten- und Evidenzmatrix dokumentieren, ohne einen künstlichen Generalisierungsscore zu erzeugen.

Modellupdates bilden dabei eine besondere Grenze. Werden Daten vor und nach einem Update zusammengefasst, beschreibt die resultierende Quote möglicherweise keine einzelne freigegebene Konfiguration. Ein Vergleich benötigt entweder getrennte Auswertung oder eine nachvollziehbare Begründung, warum die Änderung für die betrachtete Eigenschaft unerheblich ist. Dass die Benutzeroberfläche gleich aussieht, ist dafür kein ausreichendes Argument.

11. Fehlende Daten, selektive Veröffentlichung und Widerlegungsversuche

Nicht beobachtete Ergebnisse werden nicht stillschweigend als Erfolg behandelt. Ein Logausfall kann gerade mit einer Störung zusammenfallen und daher nicht zufällig sein. Die Auswertung dokumentiert fehlende Zeiträume und prüft Sensitivität: Welche Schlussfolgerung ergäbe sich unter günstigen und ungünstigen plausiblen Annahmen? Wenn der Beschluss davon abhängt, ist die Datenlücke entscheidungsrelevant.

Ebenso wichtig ist eine Suche nach Gegenbelegen. Das Gremium fragt nicht nur, welche Daten die Behauptung stützen, sondern welche Beobachtung sie widerlegen würde und ob entsprechende Fälle überhaupt in der Prüfung vorkamen. Eine Aussage über autonome Wiederherstellung ist schwach, wenn im Test keine relevanten Störungen auftraten. Ein gezielter Fehlertest darf jedoch nur unter freigegebenen Bedingungen durchgeführt werden.

Vertrauliche Daten stellen eine andere Situation dar als nicht erhobene Daten. Ein autorisierter Prüfer kann möglicherweise Einsicht erhalten, ohne alles öffentlich zu machen. Im Bericht wird dann der geprüfte Umfang genannt, während vertrauliche Einzelheiten geschützt bleiben. Unzulässig wäre die Behauptung „unabhängig bestätigt“, wenn tatsächlich nur eine Zusammenfassung des Herstellers übernommen wurde.

12. Von Messdaten zu einem belastbaren Nachweisargument

Ein Nachweisargument verbindet eine klar begrenzte Behauptung mit Unterbehauptungen, Evidenz und den Annahmen, unter denen die Verknüpfung gilt. Es kann beispielsweise erläutern, warum eine bestimmte Aufgabenklasse am Standort freigegeben werden soll. Dazu gehören nicht nur Erfolgsraten, sondern auch Schutzmaßnahmen, Fehlerbehandlung, Personenkompetenz und Konfigurationskontrolle.

Jede Verknüpfung ist angreifbar zu machen: Was würde sie ungültig machen? Eine neue Objektklasse kann den Leistungsnachweis schwächen. Eine geänderte Netzarchitektur kann die Störungsannahmen verändern. Eine nicht mehr verfügbare Serviceperson kann den Wiederanlaufprozess entwerten. Das Nachweisargument ist deshalb kein statisches Zertifikat, sondern eine versionierte Entscheidungsgrundlage.

RIG beschreibt die Verknüpfung von Entwicklung, Integration und Evidenz als Assurance-Aufgabe. Der vorliegende Leitfaden greift diese Richtung auf, beansprucht aber nicht, damit eine anerkannte Konformitätsmethode oder eine vollständige funktionale Sicherheitsanalyse bereitzustellen. [Q009, Assurance approach]

13. Verfahren eines fachlichen Prüfungsgremiums

13.1 Rollen und Vorbereitung

Vor der Sitzung werden Gegenstand, Kompetenzbereich und Interessenkonflikte festgehalten. Die Projektvertretung stellt die Unterlagen bereit, entscheidet aber nicht allein über ihre Eignung. Das Gremium sollte technische, methodische und anwendungsbezogene Kompetenz abdecken; zusätzliche Rechts- oder Sicherheitsfragen benötigen die jeweils zuständigen Fachrollen. Fehlende Kompetenz wird nicht durch Mehrheitsbeschluss ersetzt.

Die Prüfung beginnt mit der beantragten Nutzung, nicht mit einer Gesamtschau aller Projekterfolge. Danach folgen Vollständigkeit der Konfiguration, Quellenherkunft, Messdefinition, Datenintegrität, Auswertung und harte Voraussetzungen. Erst wenn diese Grundlage verstanden ist, wird über Nutzen und nächste Schritte entschieden. Eine umfangreiche Präsentation ohne zugängliche Datenakte genügt dafür nicht.

13.2 Beschluss und Nachverfolgung

Der Beschluss nennt den erlaubten Nutzungskorridor, die verantwortlichen Rollen, Bedingungen, Ablaufdatum und erneute Prüfanlässe. Mögliche Dispositionen sind gebundene Freigabe, Halt mit Nachforderung, Abbruch beziehungsweise Umsteuerung oder begründet nicht anwendbar. Offene schwerwiegende Voraussetzungen werden nicht durch einen guten Durchschnittsscore ausgeglichen.

Eine Minderheitsmeinung oder ein methodischer Vorbehalt wird dokumentiert, wenn er für die Entscheidung relevant bleibt. Das ist keine Schwäche des Gremiums, sondern Teil der Nachvollziehbarkeit. Ebenso wird festgehalten, welche neuen Informationen den Beschluss ändern würden. Eine Freigabe ohne Bedingungen und Überprüfungskriterien kann bei lernenden oder häufig veränderten Systemen schnell ihren Bezug zur Realität verlieren.

14. Reproduzierbarkeit des Prüfprozesses

Zum Datenpaket gehören Rohdaten oder ein rechtmäßig zugänglicher Prüfbestand, Datenwörterbuch, Auswertungsprogramm, Parameter, Versionen und Ergebnisdateien. Prüfsummen können nachträgliche Veränderungen erkennbar machen. Sie beweisen nicht, dass die ursprünglichen Daten wahr, vollständig oder repräsentativ sind. Eine Datei mit gültigem Hash kann einen schlecht geplanten Versuch exakt konservieren.

Die synthetischen Beispiele dieses Pakets lassen sich mit dem beigefügten Python-Skript ohne Internet und ohne zusätzliche Bibliotheken auswerten. Das Skript dient der Transparenz der Rechenwege. Es wurde nicht als sicherheitszertifizierte oder regulatorisch validierte Statistiksoftware entwickelt. Für reale hochkritische Anwendungen sind Methode, Implementierung und Datenerfassung entsprechend dem konkreten Prüfzweck unabhängig zu kontrollieren.

Prüffrage zum Abschluss: Ist unsere Schlussfolgerung enger als oder gleich weit wie die tragfähige Evidenz – oder haben wir an irgendeiner Stelle aus einer präzisen Zahl eine größere Aussage gemacht, als das Versuchsdesign erlaubt?

<!-- BEGIN AUTO-BIBLIOGRAPHY -->

Quellen zu diesem Dokument

Stabile Kennungen und genaue Fundstellen bleiben bei der PDF-Erstellung erhalten. Die genannten Quellen sind nicht sämtlich offline im Paket enthalten. Zugriffsgrenzen und Einschränkungen sind Teil des Nachweises.

Q009 · Robotics Institute Germany

Safety, Reliability and Resilience of AI-based Robotics. Stand/Publication: o. D.; Webseitenstand am Abrufdatum. Zugriff: 2026-09-25. **Original:**

<https://robotics-institute-germany.de/project/safety-reliability-and-resilience-of-ai-based-robotics/> **Fundstelle:**

Assurance approach / Research objectives **Aussagebezug:** Verknüpfung von Fähigkeitsentwicklung,

Systemintegration und Nachweiserzeugung. **Zugriff:** Relevanter Inhalt online eingesehen **Grenze:**

Forschungsprogramm; kein abgeschlossener normativer Prüfstandard.

Q015 · Figure AI

Production at BMW. Stand/Publication: 19.11.2025. Zugriff: 2026-09-25. **Original:**

<https://www.figure.ai/news/production-at-bmw> **Fundstelle:** Opening deployment account; Core KPIs; Hardware

Reliability / Forearm **Aussagebezug:** Herstellerbericht zu Figure 02: 11 Monate, 1.250+ Betriebsstunden; null

Interventionen als Ziel, nicht veröffentlichter Ist-Wert. **Zugriff:** Relevanter Inhalt online eingesehen **Grenze:**

Herstellerangabe; kein unabhängiger Audit; Definitionen der Betriebsstunden und vollständige

Interventionsverteilung fehlen.

Q016 · BMW Group

Humanoid robots in production: Leipzig / retrospective Spartanburg. Stand/Publication: 21.09.2026. Zugriff:

2026-09-25. **Original:** <https://www.bmwgroup.com/en/news/general/2026/humanoid-robot-in-leipzig.html>

Fundstelle: Spartanburg retrospective, Figure 02 **Aussagebezug:** Betreiberbericht zu Figure 02 mit 10 Monaten,

rund 1.250 Stunden und mehr als 90.000 Komponenten. **Zugriff:** Relevanter Inhalt online eingesehen **Grenze:** BMW

berichtet aus eigener Anwendung; kein unabhängiger Prüfer. Leipzig/AEON und Spartanburg/Figure nicht

verwechseln.

Q018 · NIST/SEMATECH

e-Handbook: Constant repair rate (HPP/exponential) model. Stand/Publication: o. D.; Webseitenstand am Abrufdatum. Zugriff: 2026-09-25. **Original:** <https://www.itl.nist.gov/div898/handbook/apr/section4/apr451.htm>

Fundstelle: Abschnitt 8.4.5.1; assumptions, estimates, confidence bounds **Aussagebezug:** Methodischer Bezug für

zeitbezogene Zuverlässigkeitsmodelle unter expliziten Annahmen. **Zugriff:** Relevanter Inhalt online eingesehen

Grenze: Homogener Poisson-Prozess nicht ungeprüft auf lernende oder alternde Flotten übertragbar.

12 Rechtliche Hinweise

12.1 Keine Beratung

Diese Veröffentlichung ist ein redaktionelles Erzeugnis. Sie stellt keine Rechtsberatung, keine medizinische, pflegerische, arbeitsrechtliche oder datenschutzrechtliche Beratung und keine Beschaffungs- oder Vergabeberatung dar. Sie ersetzt insbesondere nicht die Prüfung durch Rechtsanwaltschaft, Datenschutzaufsicht, Medizinprodukteberatung, Vergabestelle, Ethikkommission nach Medizinprodukterecht oder Arbeitnehmervertretung. Wer auf Grundlage dieses Dokuments entscheidet, entscheidet auf eigene Verantwortung und hat den Sachverhalt vorher eigenständig fachlich prüfen zu lassen.

12.2 Keine Zusicherung, keine Konformitätsaussage

Die beschriebenen Verfahren, Prüfstufen und Checklisten sind redaktionelle Vorschläge. Sie sind keine Zusicherung einer Eigenschaft, kein garantiertes Ergebnis und kein Nachweis der Einhaltung einer Rechtsvorschrift. Wer ein Verfahren aus diesem Dokument anwendet, erfüllt damit keine gesetzliche Anforderung; ob eine Anforderung erfüllt ist, entscheidet die zuständige Stelle am Einzelfall.

12.3 Rechtsstand

Rechtsstände, Fristen und Fundstellen beziehen sich auf den im Dokumentkopf genannten Tag. Der KI-Rechtsrahmen wird laufend geändert; ein älterer Fristenplan darf nicht unverändert weiterverwendet werden. Genannte Gesetze und Verordnungen sind in ihrer jeweils geltenden Fassung heranzuziehen.

12.4 Haftung

Die Inhalte wurden mit Sorgfalt erstellt und, soweit im Methodenteil beschrieben, gegengeprüft. Für Richtigkeit, Vollständigkeit und Aktualität wird gleichwohl keine Gewähr übernommen. Eine Haftung des Herausgebers ist auf Vorsatz und grobe Fahrlässigkeit beschränkt; dies gilt nicht bei der Verletzung des Lebens, des Körpers oder der Gesundheit sowie bei zwingender gesetzlicher Haftung.

12.5 Verweise auf fremde Inhalte

Dieses Dokument verweist auf Inhalte Dritter, insbesondere auf Behörden-, Fachgesellschafts-, Normen- und Fachverlagsseiten. Auf deren Gestaltung und Inhalt hat der Herausgeber keinen Einfluss; für sie ist jeweils der dortige Anbieter verantwortlich. Die Adressen waren zum im Literaturverzeichnis genannten Prüfzeitpunkt erreichbar. Die Nennung eines Produkts, Herstellers oder Dienstes ist weder eine Empfehlung noch eine Bewertung seiner Eignung.

12.6 Marken und Produktbezeichnungen

Genannte Produkt-, Marken- und Unternehmensbezeichnungen sind Kennzeichen ihrer jeweiligen Inhaber und werden ausschließlich zur Beschreibung des Sachverhalts verwendet. Aus der Nennung folgt keine Verbindung zwischen dem Herausgeber und dem jeweiligen Unternehmen und keine Aussage über Zulassungs-, Konformitäts- oder Zertifizierungsstatus.

12.7 Urheberrecht und Nutzung

Dieses Dokument, sein Text, seine Tabellen und seine Abbildungen sind urheberrechtlich geschützt. Rechteinhaberin ist die WebTrades GmbH. Zulässig sind das Lesen, das Speichern und der Ausdruck für den eigenen Gebrauch sowie das Zitieren einzelner Stellen mit vollständiger Quellenangabe nach der im Impressum genannten Zitierweise. Die Vervielfältigung, Bearbeitung, Verbreitung, öffentliche Zugänglichmachung oder kommerzielle Verwertung des Dokuments oder wesentlicher Teile daraus bedarf der vorherigen schriftlichen Zustimmung der Rechteinhaberin. Das gilt ausdrücklich auch für die Übernahme von Tabellen und Abbildungen.

12.8 Keine personenbezogenen Daten

Für dieses Dokument wurden keine personenbezogenen Daten verarbeitet. Es beruht auf veröffentlichter Literatur, amtlichen Rechtsquellen und Veröffentlichungen von Fachgesellschaften; es wurden weder Beschäftigte befragt noch Patientendaten verwendet. Wer die hier beschriebenen Verfahren auf eine reale Einrichtung anwendet, verlässt diesen Rahmen: Die Auswertung von Tätigkeits-, Alarm- oder Quittierungsdaten unterliegt der Datenschutz-Grundverordnung und in der Regel der Mitbestimmung.

12.9 Unabhängigkeit

Dieses Dokument ist ohne Auftrag, ohne Finanzierung und ohne inhaltliche Beteiligung eines Herstellers, Krankenhausträgers, Kostenträgers, Verbands oder Beratungsunternehmens entstanden. Es besteht kein Vergütungs- oder Beteiligungsverhältnis zu einem der genannten Unternehmen oder Institutionen. Genannt werden Systeme, Anbieter und Dokumente, weil sie in den beschriebenen Gegenstand fallen, nicht weil sie beworben werden sollen.

12.10 Anwendbares Recht

Es gilt das Recht der Bundesrepublik Deutschland unter Ausschluss des UN-Kaufrechts. Gesetzliche Verbraucherschutzvorschriften bleiben unberührt.